Title: Appendix for "This photograph has been altered: Testing the effectiveness of image forensic labeling on news image credibility"
Authors: Cuihua Shen (1), Mona Kasra (2), James F. O'Brien (3)
Date: May 25th, 2021
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix

Experimental design: Additional factors

Image credibility cues (not hypothesized). The "high credibility" condition was achieved in using New York Times as the purported source, with high virality metrics. The "low credibility" condition was achieved by using a generic person's Twitter account (Rachael Hughes) as the purported source of the image, with very low virality metrics. The image sources were selected based on a Pew report on media trustworthiness (Mitchell et al., 2014), which ranked the New York Times as one of the most trustworthy news sources. Both the purported sources and virality metrics were validated and used in a previous study (Shen et al., 2019).

Image content (not hypothesized). These images were used in a previous study (Shen et al, 2019) and represented different sociopolitical issues with varied media exposure in recent years.

These two factors were included to expand ecological validity of the study, not to test their separate effects on the outcome variable, so they were manipulated but not explicitly tested in the analysis.

Measures

Perceived credibility. This variable was measured by six items of perceived credibility adapted from Flanagin and Metzger's (2007) on a 7-point scale (1 = strongly disagree, 7 = strongly agree). It assessed the extent to which participants perceived the image to be believable, original, authentic, fake, manipulated, and retouched. After reverse-coding negatively-worded items, the mean was taken to create a composite credibility score (*Cronbach's alpha* = .95). In the concurrent exposure and control conditions, credibility was measured only once. In the post-exposure condition, the perceived credibility was measured twice, once before seeing the barometer and once after. We also calculated the net credibility change by subtracting the pre-barometer rating from the post-barometer rating.

Internet skills. Participants' Internet skills were measured by their familiarity with 10 Internet-related terms (e.g. phishing and spyware) on a 5-point Likert scale (Hargittai & Hsieh, 2012). Then, the mean of these items became a composite score of Internet skills (*Cronbach's alpha* =.92).

Digital imaging skills. Two items were used to measure participants' photography and digital imaging (e.g. photo editing) experiences and skills (Greer & Gosen, 2002) on a 5-point scale (1 = None, 5 = I'm an expert). The mean was then taken to be the composite measure of digital imaging skills (*Cronbach's alpha* = .74).

Pro-issue attitude. For each of the three images tested, two items were used to measure participants' preexisting attitudes toward the issue depicted in the image. These items were adapted from Treier and Hillygus (2009) and modified to fit each of the images tested. For example, participants evaluating the image showing a genetically modified mouse were asked whether it is ethical or acceptable to genetically modify animals for research purposes. Negatively worded questions were reversed coded, and then the two items were averaged to create a composite score of pro-issue attitude (*Cronbach's alpha*=.81).

Political ideology. Participants were asked to indicate their political ideology on a 7-point Likert scale, from extremely conservative (1) to extremely liberal (7).

Internet use. Participants were asked how many years they have been using the Internet, and also how many hours on average per day they use the Internet for non-work reasons.

Demographics. At the end of the survey, participants were asked to indicate their sex, age, race,

annual household income, and education level. Participants' age and sex were included in our analysis as control variables.

Manipulation check

The study performed a manipulation check of the forensic label by asking participants to indicate what forensic designation the barometer was pointing at (Un-altered, Altered, or not sure). Among the 2440 participants who completed the study, 2283 were exposed to an image forensic barometer (conditions 1-12), of which 1982 (86.8%) correctly recalled its forensic designation, and 301 (13.2%) either answered the wrong designation or "unsure.". As expected, those who failed the manipulation check rated the image more credible than those who identified the forensic designation correctly (M_{failed} = 3.39, M_{passed} = 3.13, t = -2.49, p = 0.01). A chi-square test showed that participants in the post-exposure placement condition were more likely to fail the manipulation check than those assigned to the concurrent placement condition (*Chi-square* = 37.335, *df* = 1, p <.001). In the following analysis, these participants who failed the manipulation check (n = 301) were excluded, leaving a final sample of 2139 participants.

Findings 1 & 2 analysis

To test the main effect of image forensic labeling, we ran analyses in two stages. In the first stage, an omnibus ANOVA showed a significant main effect (F(2, 2136)=112.38, p<.001). Multiple comparisons using Dunnett T3 and Bonferroni adjustment showed that participants exposed to the "Altered" label rated the image significantly less credible than those who did not see the label (M_{diff} = -1.16, p < .001), but those exposed to the "Un-altered" label did not differ from the control group (M_{diff} = 0.03, p = .99). In the second stage, we ran an analysis of covariance (ANCOVA) with perceived credibility of the image (the second credibility rating for post-exposure condition) as the dependent variable, and image forensic label as the main factor, while also including the respondent's age, gender, political ideology, and issue attitude as covariates. Results still showed a significant main effect of image forensic label (F(2, 2126) = 120.96, p < .001). A planned contrast between the "Un-altered" condition and the control condition showed a non-significant difference (M_{diff} = -0.129, SE = 0.15, p=.40), while participants in the the "Altered" condition rated image significantly less credible than those in the control condition (M_{diff} = -1.31, SE = 0.15, p < .001). Therefore, both ANOVA and ANCOVA showed the same results.

Among the covariates, people's prior experience with digital imaging and photography has a significant and negative association with credibility ratings (F(1, 2126) = 7.86, p = .005, B = -0.15), so did people's internet skills (F(1, 2126) = 7.72, p = .005, B = -0.14). Their pre-existing attitude supporting the issue depicted in the image showed a significant and positive association with credibility (F(1, 2126) = 86.45, p < .001, B = 0.20). Participant's pre-existing political affiliation (F = 3.62, p = .06), age (F(1, 2126) = 0.93, p = .34), and gender (F(1, 2326) = 0.58, p = .45) did not associate with how they rated credibility of these images.

To probe whether the results differed across the two image credibility cues conditions (high vs. low credibility cues), we ran a post-hoc ANCOVA with image credibility cues as an additional factor, along with its interaction term with image forensic designation. We found that both the main effect of credibility cues (F(1, 2123) = 2.20, p =.14) and the interaction between credibility cues and forensic labels (F(1, 2123) = 1.446, p =.229) were nonsignificant. Therefore, our results are robust across different credibility cue manipulations.

To probe whether demographic groups differ in digital media literacy, we ran omnibus tests between male and female respondents. We found that, compared to men, women in our sample have slightly lower internet skills (M_{men} = 4.15, M_{women} = 3.93, t = -6.20, df = 2137, p <.001) as well as lower digital imaging experiences (M_{men} = 2.89, M_{women} = 2.75, t= -3.92, df =2137, p <.001). But no significant differences in digital media literacy exist among racial groups (Internet skills: F(7, 2131) = 0.48, p = 0.85; Digital imaging experiences: F(7, 2131) = 1.82, p = 0.08).

Finding 3 analysis

In order to test whether participants' exposure to visual misinformation would have a continued influence effect after they were shown the forensic label, we ran paired sample t-tests between their first credibility rating (before seeing the forensic label) and their second credibility rating of the same image (after seeing the forensic label). Their second rating was significantly higher than the first rating for those who saw the "Un-altered" label ($M_{difference}$ = 0.33, t = 6.88, p <.001), and significantly lower than the first rating for those who saw the "Altered" label ($M_{difference}$ = -0.83, t = -14.98, p <.001). Additionally, ANOVA tests showed that participants' second rating of image credibility and was statistically equivalent to those of the concurrent condition (Participants exposed to the "Altered" label: F(1, 978) = 1.96, p = .16; Participants exposed to the "Un-altered'' label: F(1, 1000) = 0.39, p =.53). To test whether the results were robust across different image credibility cue conditions and with covariates, we ran ANCOVA models with image credibility cues as an additional factor. Results were virtually unchanged, and no significant difference was found across the high and low credibility cue conditions.

Finding 4 analysis

To test the main effects of labeling source, we ran two sets of models, one with the participants who were shown the "Altered" label, and the other with participants shown the "Un-altered" label. The omnibus ANOVA test showed that the source of forensic label with three levels (experts, other people online, and software) on its own was not associated with participants' credibility perception of the images (Participants exposed to the "Altered" label: F(2, 977) = 2.25, p = .11; Participants exposed to the "Unaltered" label: F(2, 999) = 0.44, p = .64).

Post-hoc two-way ANOVA of both source and placement of forensic labels showed that the interaction between rating source and placement was significant for participants exposed to the "Altered" label (F(1, 974) = 3.31, p = .04) but not for participants exposed to "Un-altered" label (F(2, 996) = 1.70, p = .18). Specifically, if the "Altered" label's forensic analysis purportedly came from software instead of experts or other people online, its association with people's credibility perception bifurcated in post exposure and concurrent conditions (see Figure 7).

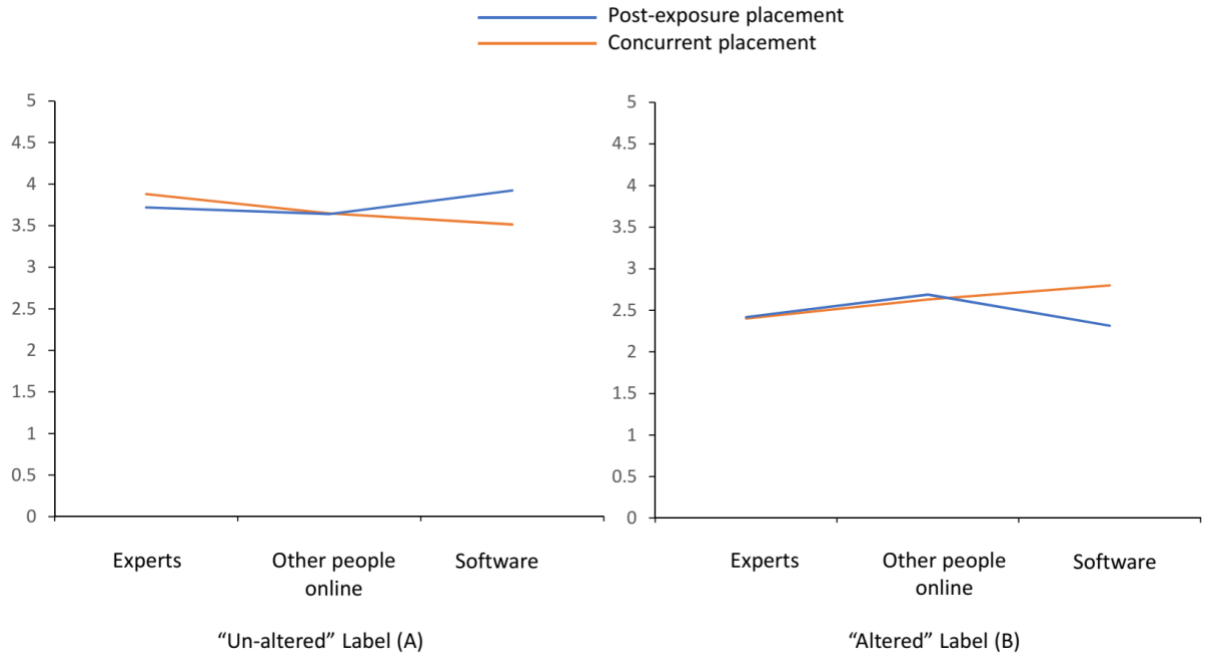

Figure 7. Interactions between forensic label placement and source of forensic label. Higher numbers represent greater perceived image credibility. *Panel (A) shows the participants exposed to the "Un-altered" label, and Panel (B) shows those exposed to the "Altered" label.*

Based on the above results, we consolidated the source of forensic analysis into just two levels: "human" (combining "expert" and "other people online") versus "software." We then ran two-way ANOVA with the source and placement of forensic labels. Results again showed a significant interaction for those seeing the "Altered" label ($F(1, 976) = 6.67$, $p = .01$), and a marginally significant interaction for those seeing the "Un-altered" label ($F(1, 998) = 3.25$, $p = .07$).

To test whether the results differed across the two image credibility cues conditions (high vs. low credibility cues), we further added image credibility cues as another factor along with covariates. Main results are unchanged from previous models, and the three-way interaction among credibility cues, rating source and placement of forensic analysis was not significant (participants seeing the "Altered" label: $F(1,965) = 0.29$, $p = .59$; participants seeing the "Un-altered" label: $F(1,985) = 0.60$, $p = .43$), showing that the results were robust and did not differ across high and low image credibility cue conditions.

Limitations and future research

Our study has a number of limitations that can be explored in future research. First, our forensic barometer, Picture-O-Meter, followed the design of Politifact's Truth-O-Meter, which included three midpoints in addition to the two extremes. However, our barometer did not have textual indicators for these three midpoints, which might have confused participants. In real-world deployment of forensic labels, a machine learning classifier would produce an estimated probability that the image is or is not altered. Future research needs to explore if showing intermediate values and indicating uncertainty makes the labels more or less credible. Second, our study only examined news images, while excluding cartoons, infographics, and memes. Further research should examine how image forensic labels may help counteract misinformation presented in infographics and memes, which are prevalent online and may be considered as credible despite their synthetic nature. Third, our study tested only three images covering three socio-political issues. Even though the findings were robust across all three images, the

generalizability of our findings, therefore, needs to be further assessed using a larger pool of images and issues. Lastly, our study utilized an artificial experimental setting that was different from participants' real-world news consumption context. We were unable to measure actual user engagement behaviors such as liking, commenting and sharing of the news posts. We also relied upon a Mturk sample, which was compensated less than the minimum wage and may deviate from the general population. Future research is encouraged to use actual news platforms and more representative samples to verify the real-world validity of our findings.